DPIR SPRING SCHOOL

Turning text into data: some programming basics

11 April, 2016

Prof. Andrew Eggers

R and regular expressions: the big picture

For someone who wants to use R for text analysis, what is most important?

(1) Knowing **how to find answers** when you have questions

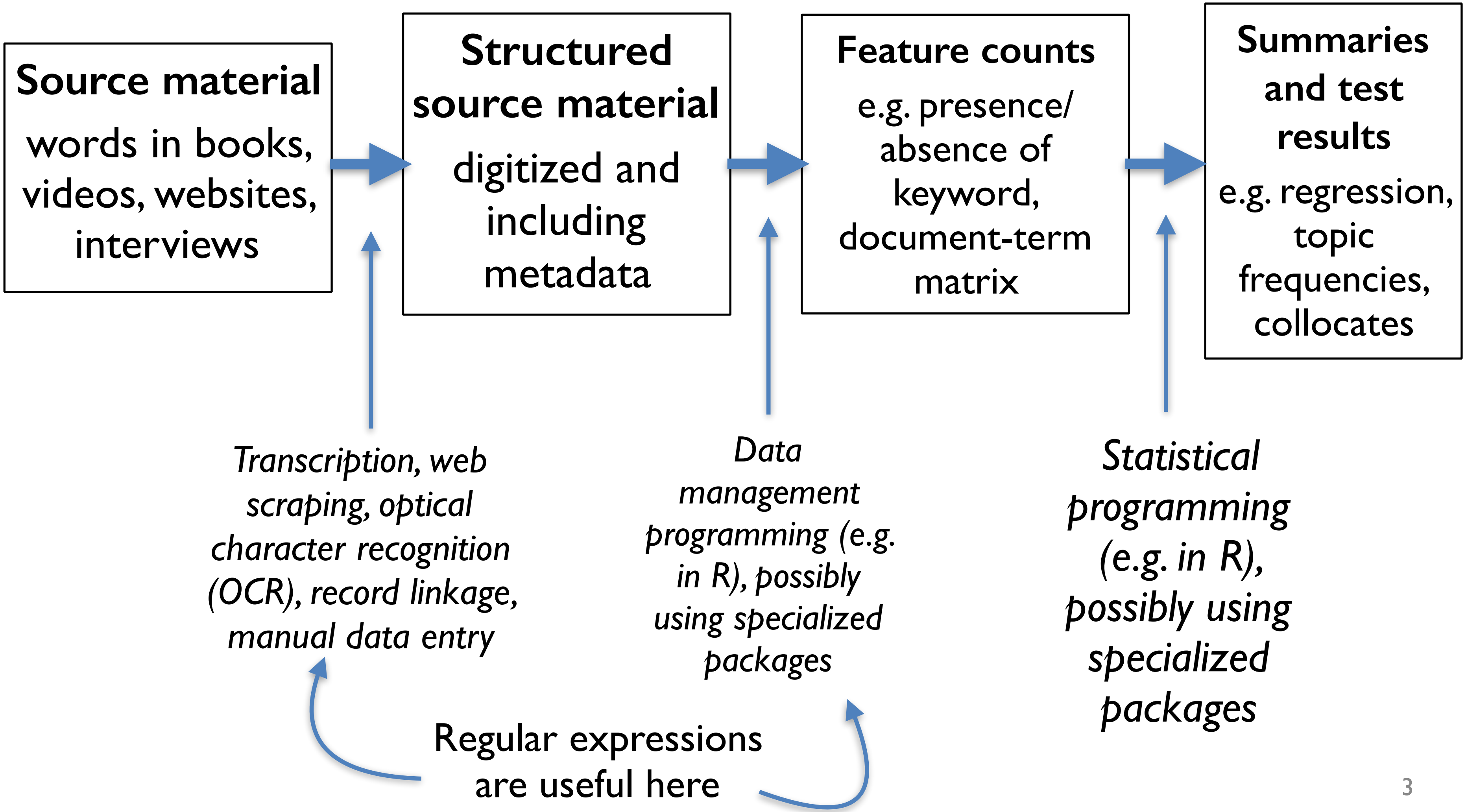
(2) Understanding the **basic syntax/structure** of R: data types, data structures, input/output

(3) Knowing how to do **basic operations on character strings**:

- *regular expressions* for find/replace, matching, counting
- other basic stuff: splitting, joining, displaying, etc

(4) Finding and using **package(s)** that contain the most important utilities for transforming and analyzing text data

A workflow for text analysis



Source material

words in books, videos, websites, interviews

Structured source material

digitized and including metadata

Feature counts

e.g. presence/ absence of keyword, document-term matrix

Summaries and test results

e.g. regression, topic frequencies, collocates

Transcription, web scraping, optical character recognition (OCR), record linkage, manual data entry

Data management programming (e.g. in R), possibly using specialized packages

Statistical programming (e.g. in R), possibly using specialized packages

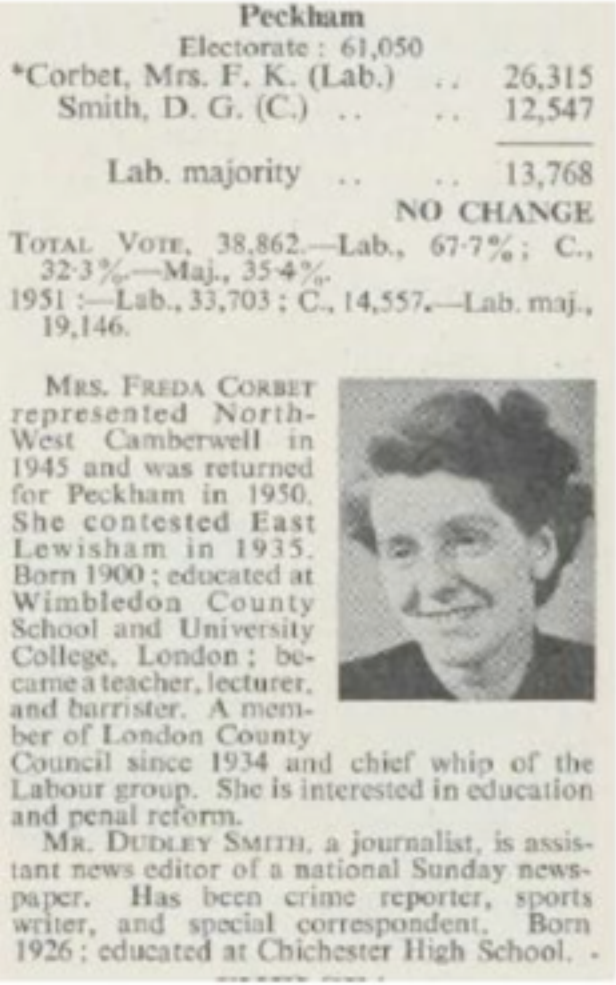
Regular expressions are useful here

What do we do with regular expressions?

- Identifying/counting mentions of specific words/phrases:
  - Given open-ended responses on a survey, identify the ones that mention "economy" "jobs" "unemployment" etc
  - Given a dataset of parliamentary speeches, count mentions of "Ireland" or "Irish" in each speech
  - Find the candidate biographies that mention Oxbridge, Eton, etc
- Collection and pre-processing of the corpus:
  - Web scraping, correcting OCR errors, removing HTML tags, identifying and extracting relevant content (e.g. speeches, not motions)
  - For sentiment analysis, converting "hardly a good idea" to "not good" before counting

Example: Eggers and Hainmueller (2009) "MPs for Sale?"

7 volumes of *Times Guide to the House of Commons*

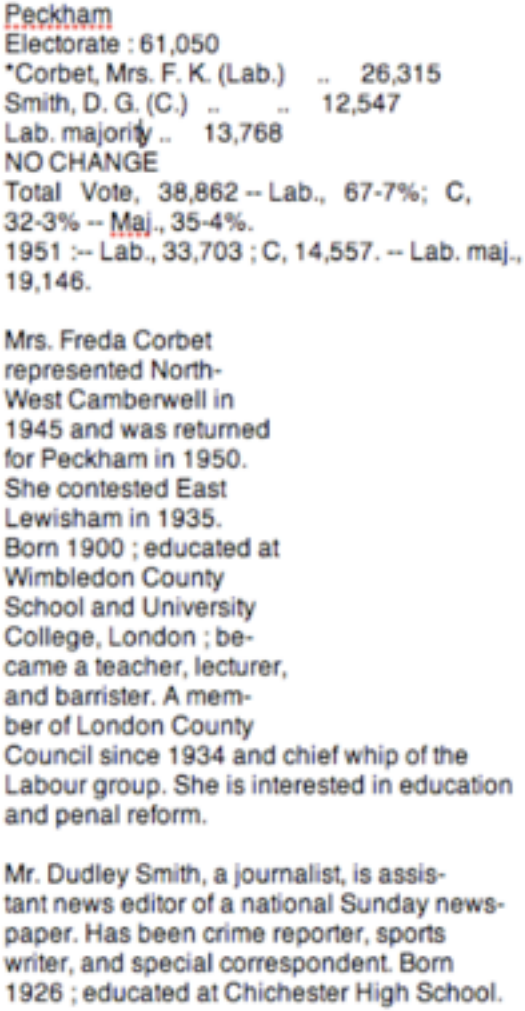


Peckham
Electorate : 61,050
*Corbet, Mrs. F. K. (Lab.) .. 26,315
Smith, D. G. (C.) 12,547
Lab. majority 13,768
NO CHANGE
TOTAL VOTE, 38,862.—Lab., 67·7%; C., 32·3%.—Maj., 35·4%.
1951 :—Lab., 33,703 ; C., 14,557.—Lab. maj., 19,146.

MRS. FREDA CORBET represented North-West Camberwell in 1945 and was returned for Peckham in 1950. She contested East Lewisham in 1935. Born 1900 ; educated at Wimbledon County School and University College, London ; became a teacher, lecturer, and barrister. A member of London County Council since 1934 and chief whip of the Labour group. She is interested in education and penal reform.

MR. DUDLEY SMITH, a journalist, is assistant news editor of a national Sunday newspaper. Has been crime reporter, sports writer, and special correspondent. Born 1926 ; educated at Chichester High School.

Converted to text by Widener Library digital services



Peckham
Electorate : 61,050
*Corbet, Mrs. F. K. (Lab.) .. 26,315
Smith, D. G. (C.) 12,547
Lab. majority .. 13,768
NO CHANGE
Total Vote, 38,862 -- Lab., 67-7%; C., 32-3% -- Maj., 35-4%.
1951 :-- Lab., 33,703 ; C, 14,557. -- Lab. maj., 19,146.

Mrs. Freda Corbet represented North-West Camberwell in 1945 and was returned for Peckham in 1950. She contested East Lewisham in 1935. Born 1900 ; educated at Wimbledon County School and University College, London ; became a teacher, lecturer, and barrister. A member of London County Council since 1934 and chief whip of the Labour group. She is interested in education and penal reform.

Mr. Dudley Smith, a journalist, is assistant news editor of a national Sunday newspaper. Has been crime reporter, sports writer, and special correspondent. Born 1926 ; educated at Chichester High School.

Converted to database using **regular expressions** to identify party, vote count, profession, school, date of birth for each candidate

Introduction and practice with regular expressions

(To be done in RStudio.)

Principles of coding that apply here

- You need to practice/use it, ideally with a real problem!
- All of your work needs to be in a well-commented script, which takes raw data to finished product
- You need to test what you've done; at the very least check often that you get what you would expect

And specifically about regular expressions

- Most characters just match themselves in regular expression (e.g. `z f e`)
- Meta-characters have a special meaning, e.g. `. ( ) ^ $ * ? + [ ]`
- To actually search a dollar-sign or a plus-sign: `\\$ \\+`
- Other special characters:
  - `\\d  \\s` etc
  - `[[:digit:]] [[:punct:]]`